REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 26-04-2011	**2. REPORT TYPE** Article	**3. DATES COVERED** *(From - To)* APR 2011 - MAY 2011

4. TITLE AND SUBTITLE
Phonologically-Based Biomarkers for Major Depressive Disorder

5a. CONTRACT NUMBER
FA8720-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Andrea C. Trevino, Thomas F. Quatieri, and Nicolas Malyska

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ASD R&E

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Of increasing importance in the civilian and military population is the recognition of Major Depressive Disorder at its earliest stages and intervention before the onset of severe symptoms. Toward the goal of more effective monitoring of depression severity, we introduce vocal biomarkers that are derived automatically from phonologically-based measures of speech rate. To assess our measures, we use a 35-speaker free-response speech database of subjects treated for depression over a six-week duration. We find that dissecting average measures of speech rate into phone-specific characteristics and, in particular, combined phone-duration measures uncovers stronger relationships between speech rate and depression severity than global measures previously reported for a speech-rate biomarker. Results are supported by correlation of our measures with depression severity and classification of depression state with these vocal measures. Our approach provides a general framework for analyzing individual symptom categories through phonological units, and supports the premise that speaking rate can be an indicator of psychomotor retardation severity.

15. SUBJECT TERMS
Major Depressive Disorder, vocal biomarkers, speech rate, speech, phone, clinical HAMD

16. SECURITY CLASSIFICATION OF: U			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Zach Sweet
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	15	**19b. TELEPHONE NUMBER** *(include area code)* 781-981-5997

Phonologically-Based Biomarkers for Major Depressive Disorder[1]

Andrea C. Trevino, Thomas F. Quatieri, and Nicolas Malyska

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

Abstract:

Of increasing importance in the civilian and military population is the recognition of Major Depressive Disorder at its earliest stages and intervention before the onset of severe symptoms. Toward the goal of more effective monitoring of depression severity, we introduce vocal biomarkers that are derived automatically from phonologically-based measures of speech rate. To assess our measures, we use a 35-speaker free-response speech database of subjects treated for depression over a six-week duration. We find that dissecting average measures of speech rate into *phone-specific characteristics* and, in particular, *combined phone-duration measures* uncovers stronger relationships between speech rate and depression severity than global measures previously reported for a speech-rate biomarker. Results are supported by correlation of our measures with depression severity and classification of depression state with these vocal measures. Our approach provides a general framework for analyzing *individual symptom* categories through phonological units, and supports the premise that speaking rate can be an indicator of psychomotor retardation severity.

Index Terms—**Major Depressive Disorder, vocal biomarkers, speech rate, speech, phone, clinical HAMD.**

[1] This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

I. Introduction

Major Depressive Disorder (MDD) is the most widely affecting of the mood disorders; the lifetime risk has been observed to fall between 10-20% for women and 5-12% for men [6]. In addition, the 2001 World Health Report names MDD as the most common mental disorder leading to suicide [1][26]. Currently, no laboratory markers have been determined for the diagnosis of MDD, although a number of abnormalities have been observed when comparing patients with depression to a control group [1]. Accurate diagnosis of MDD requires intensive training and experience, thus the growing global burden of depression suggests that an automatic means to help detect and/or monitor depression would be highly beneficial to both patients and healthcare providers. One such approach relies on the extraction of biomarkers to provide reliable indicators of depression.

One class of biomarkers of growing interest is the large group of vocal features that have been observed to change with a patient's mental condition and emotional state. Examples include vocal characteristics of prosody (e.g., pitch and speech rate), spectral features, and glottal (vocal fold) excitation patterns [2][8][12][13][15][16][18][19]. These vocal features have been shown to have statistical relationships with presence and severity of certain mental conditions, and, in some cases, have been applied towards developing automatic classifiers. In this paper, we expand on work for the particular prosodic biomarker of *speech rate*, which has been shown to significantly separate control and depressed patient groups [25]. Specifically, we present vocal biomarkers for depression severity derived from *phonologically-based measures of speech rate*. In addition, we investigate this dependence with respect to each of the *symptom-specific components* that comprise the standard 17-item HAMD [9] composite assessment of depression. For example, supporting the premise that psychomotor retardation can be observed in the speech rate [10][25], we reveal high correlations between not only the global speech rate but also a subset of individual phone durations and the HAMD Psychomotor Retardation sub-topic. Although the specific focus in this paper is biomarkers derived from speech rate, we provide a general framework in which to explore the relationship between phonologically-based biomarkers and the severity of individual MDD symptoms.

In this work, we investigate the correlations between phonologically-based biomarkers and the clinical HAMD severity ratings, for a 35-speaker free-response speech database, recorded by Mundt et al. [16]. We first compute global speech rate measures and show the relationship with the HAMD total and sub-topic ratings through correlation studies; these global rate measures are computed by finding the average phone rate using an automatic phone recognition algorithm. We then examine the correlations of the HAMD ratings with the average duration of pauses and automatic recognition-based individual English phone durations, providing a *fine-grained analysis of speech timing*. With regard to the pause measures, the findings with pause duration are consistent with previous total HAMD rating correlations [16], but extend the analysis to the sub-topics. With regard to the individual phone durations (vowels and consonants), higher individual correlation values than those found with the global speech rate measures reveal distinct phone-specific relationships. The individual phone durations that show significant correlations within a single HAMD category (total or sub-topic) are observed to cluster approximately within manner-of-articulation categories and according to the strength of intercorrelation between sub-topics. These significantly correlated phone lengths within a sub-topic are then selected and linearly combined to form composite durations; these composite durations result in correlation values that exceed those found not only using the individual phone durations but also the more global vocal measures that are used in our work and previous studies [16]. As an extension of the individual phone duration results, the energy spread of a phone is provided as an alternate duration measure; the energy spread measure reveals some similar phone-specific correlation patterns and more changes in correlations with burst consonants relative to those calculated from the recognition-based duration. A broad overview of our phonologically-based (fine-grained timing) framework with an included list of our key measures is illustrated in Figure 1.

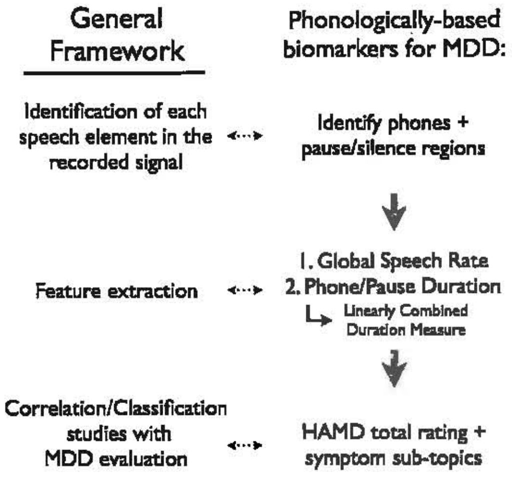
General Framework
Phonologically-based biomarkers for MDD:
Identification of each speech element in the recorded signal
Identify phones + pause/silence regions
Feature extraction
1. Global Speech Rate
2. Phone/Pause Duration
Linearly Combined Duration Measure
Correlation/Classification studies with MDD evaluation
HAMD total rating + symptom sub-topics

Figure 1: Overview of the general framework presented in this paper and our specific approach.

We conclude with a preliminary classification investigation using our phonologically-based duration measures, guided by the significant correlations from our phone-specific results. Using a simple Gaussian-likelihood classifier, we examine the accuracy in classifying the individual symptom sub-topic ratings by designing a multi-class classifier where each rating level is set as its own class. The classification root mean squared error (RMSE) is reported as a measure of accuracy. Our preliminary classification results show promise as a beneficial tool to the clinician, and motivate the addition of other phone-based features in classification of depression severity.

Our results provide the framework for a phone-specific approach in the study of vocal biomarkers for depression, as well as for analyzing individual symptom categories. To further exploit this framework, the scarcity and variability of samples in our database points to a need for further experiments with larger populations in order to account for the variety within one group of MDD patients.

II. BACKGROUND AND PREVIOUS WORKS

A. Major Depressive Disorder

Major Depressive Disorder (MDD) places a staggering global burden on society. Of all mental disorders, MDD accounts for 4.4% of the total disability-adjusted life years (DALYs)[2] lost and accounts for 11.9% of total years lost due to disability (YLD). With current trends, projection for the year 2020 is that depression will be second only to ischemic heart disease as the cause of DALYs lost worldwide [26].

B. Diagnosis and Treatment

Major Depressive Disorder (MDD) is characterized by one or more Major Depressive Episodes (MDE), where an MDE is defined as a period of at least two weeks during which either a depressed mood dominates or markedly diminished interest, also known as anhedonia, is observed. Along with this, the American Psychiatric Association standard recommends that at least four or more of the following symptoms also be present for diagnosis: significant change in weight or appetite, insomnia or hypersomnia nearly every day, psychomotor agitation or retardation (clearly observable by others), fatigue, feelings of worthlessness or excessive guilt, diminished ability to concentrate or decide, and/or recurrent thoughts of death or suicide [1]. These standards are reflected in the HAMD depression rating method, which encompasses multiple symptoms in order to gauge the overall severity of depressive state, as discussed further in the following section. Conventional methods for treatment of MDD include pharmacotherapy and/or psychotherapy; exhaustive coverage of depression treatment is beyond the scope of this paper.

C. Depression Evaluation – HAMD

We consider the standard method of evaluating levels of MDD in patients, the clinical 17-question HAMD assessment (a detailed description of the database is given in Section III). To determine the overall or total score, individual ratings are first determined for symptom sub-topics (such as mood, guilt, psychomotor retardation, suicidal tendency, etc.); the total score is then the aggregate of the ratings for all sub-topics. The sub-topic component list for the HAMD (17 symptom sub-topics) evaluation is provided in the Appendix. Scores for component sub-topics have ranges of (0-2), (0-3), or (0-4).

Although the HAMD assessment is a standard evaluation method, there are well-known concerns about its validity and reliability [2]. Nevertheless, our purpose in this paper is not to test whether the HAMD (or its sub-topic ratings) are valid, but to instead provide a flexible analysis framework that can be adapted to future depression evaluation standards. The interdependencies for our particular database are discussed in Section III.

D. Previous Works

In this section, we provide a representative sampling of vocal features previously applied as MDD discriminators through correlation measurements and/or classification algorithms. These vocal measurements fall into the broad categories of prosody (e.g., pitch and speech rate), spectral, glottal (vocal fold) excitation, and energy (power).

We begin with an early study by Flint et al. [7] who used the second formant transition, voice onset time, and spirantization, a measure that reflects aspirated "leakage" at the vocal folds, to discriminate between MDD, Parkinson's disease, and control subjects. Although significant ANOVA (analysis of variance) differences were computed for a small feature subset, no significant correlations between any of the features and the HAMD scores were found in the depression studies.

France et al. [8] later used similar biomarkers including the fundamental frequency, amplitude modulation, formant statistics, and power distribution to classify control, dysthymic, MDD, and suicidal males and females, separately. The female vocal recordings showed spectral flattening with MDD; the results for the male recordings showed that the location and bandwidth of the first formant along with the percent of total power in the 501-1000 Hz sub-band were the best discriminators between the MDD subjects and the controls.

Ozdas et al. [18][19] investigated the use of two vocal features, vocal-cord jitter and the glottal flow spectrum, for differentiating between control, MDD, and near-term suicidal risk subjects. Depressed and near-term suicidal patients showed increased vocal-cord jitter and glottal spectral slope.

Moore, in a series of papers [14][15], also investigated vocal glottal excitation, spectral, and prosodic characteristics. A large variety of statistical measures were then utilized to construct classifiers for distinguishing control from depressed patient groups; these classifiers were used to infer the most differentiating feature-statistic combinations for their dataset.

Low et al. [12] combined prosodic, spectral, and the first and second derivatives of the mel-cepstra features to classify control and clinically depressed adolescents, using a Gaussian Mixture Model (GMM)-based classifier. With a combination of these vocal features, the final classification accuracy was able to reach 77.8% and 74.7% for males and females, respectively.

A study by Mundt et al. [16] showed that depressed patients responding to treatment significantly increased their pitch variability about the fundamental frequency more than non-responders. This analysis also suggested that depressed

[2] The World Health Organization defines DALYs for a disease as the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition [28].

patients may extend their total vocalization time by slowing their syllable rate and through more frequent and longer pause times. The results of Mundt et al. provide a springboard for our current effort. In contrast to the Mundt work, which uses the assumed fixed number of syllables in the "Grandfather Passage" to analyze speech rate, our work focuses on the conversational free-response speech recordings and performs a fine-grained analysis by using automatically-detected individual phone durations. More detailed comparisons to the results of Mundt et al. are provided in the Measurements sections, where comparative measures are analyzed.

As one of the emerging approaches to depression recognition, Cohn et al. [4] aimed at fusing facial and vocal features to create a more accurate MDD classifier. Measures of vocal prosody included average fundamental frequency and participant/speaker switch duration. Using a Support Vector Machine (SVM) classifier, true positive and negative rates of 88% and 64%, respectively, were achieved from these vocal features.

Certain vocal features in MDD studies are also tracked in studies of vocal affect and emotion. Among these features are changes in mean fundamental frequency, mean intensity, and rate of articulation, as well as standard spectral-based speech analysis features such as the mel-cepstrum [5][20].

The vocal biomarker studies described in this section generally take a global approach to speech, as opposed to phone or phonological group-specific effects. In addition, these works focus primarily on the total evaluation ratings or group depressed patients into one large set, regardless of sub-symptom variability. In contrast, the approach of this paper relies on *decomposition of the speech signal into unique phones and of the total depression score into individual symptom sub-topic ratings*, thus providing a unique framework for detailed analysis of unit-dependent vocal features, and how they change with individual aspects of depression severity.

III. DATABASE

The data used in this analysis was originally collected by Mundt et al. (2006) for a depression-severity study, involving both in-clinic and telephone-response speech recordings [16]. Thirty-five physician-referred subjects (20 women and 15 men, mean age 41.8 years) participated in this study. The subjects were predominately Caucasian (88.6%), with four subjects of other descent. The subjects had all recently started on pharmacotherapy and/or psychotherapy for depression and continued treatment over a 6-week assessment period. Speech recordings (sampled at 8 kHz) were collected at weeks 0, 2, 4, and 6 during an interview and assessment process that involved HAMD scoring. To avoid telephone-channel effects, only the samples of conversational (free-response) speech recorded in the clinic are used in our follow-up work. Additionally, we only used data from subjects who completed the entire longitudinal study. This resulted in approximately 3-6 minutes of speech per session (i.e., per day). More details of the collection process are given in [16].

Ratings from the 17-item HAMD clinical MDD evaluation were chosen as comparison points in our study. Individual sub-topic ratings from each evaluation (see Appendix) were also used both in our correlation studies and classification-algorithm development.

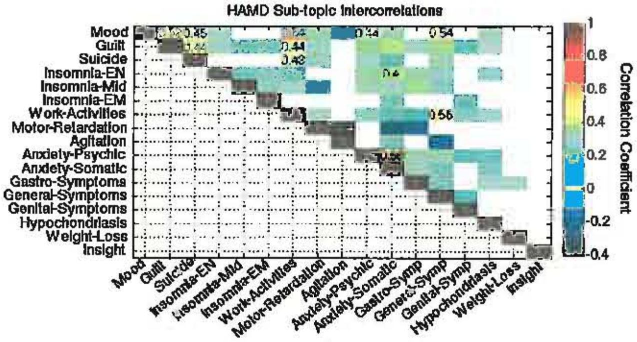

Figure 2: Table of HAMD sub-topic intercorrelations; only significant (p-values<0.05) correlations are shown with non-zero magnitude. Color bar indicates the sign and magnitude of correlation coefficient. All correlations values greater than 0.4 in absolute value are listed in the table. For clarity, all values below the diagonal of the symmetric correlation matrix have been omitted.

An important additional consideration is the intercorrelations between the HAMD symptom sub-topics. Figure 2 shows all significant intercorrelations between the HAMD sub-topics, computed with our dataset. The greatest absolute correlation of 0.64 corresponds to the Mood and Work-Activities sub-topics. High significant correlations group the sub-topics of Mood, Guilt, Suicide, and Work-Activities together. Relevant to the findings in this work, the Psychomotor Retardation sub-topic has the strongest correlations with Agitation (-0.40) and Mood (0.36, not labeled).

IV. GLOBAL RATE MEASUREMENTS

Our approach is based on the hypothesis that general psychomotor slowing manifests itself in the speech rate, motivated by observed psychomotor symptoms of depression [7][25] and supported by previous findings of correlation between MDD diagnosis and/or severity with measures of speech rate [16]. In our work, we investigate a measure of speech rate derived from the durations of individual phones. For the phone-based rate measurements, we use a phone recognition algorithm based on a Hidden Markov Model approach, which was reported as having about an 80% phone-recognition accuracy [24]. Possible implications of phone-recognition errors are discussed in Section V.

We compute the number of speech units per second over the entire duration of a single patient's free-response session. We use the term *speaking rate* to refer to the phone rate over the total session time, with times when the speech is not active (pauses) included in the total session time. This is in contrast to *articulation rate,* which is computed as the phone rate over only the time during which speech is active.

Phone rates were computed for each individual subject and session day using the database described in Section III (i.e., the in-clinic free-response speech in the collection by Mundt et. al. [16]). Correlations between these global rate measures and the total HAMD score, along with its sub-topics (17 individual symptom sub-topics), were all computed. For our results, Spearman correlation was chosen over Pearson due to the quantized ranking nature of the HAMD depression scores and the possible non-linear relationship between score and speech feature [13][17]. Thus, the correlation results determine if a monotonic relationship exists between extracted speech features and depression-rating scores.

TABLE I
SCORE CORRELATIONS WITH SPEAKING AND ARTICULATION RATE

Rate Measure	Score Category	Spearman Correlation	p-value
Speaking - Phone Rate	HAMD Work and Activities	-0.20	$0.01<p<0.05$
	HAMD Psychomotor Retardation	-0.38	$p = 3.6e-5$
	HAMD TOTAL	-0.22	$0.01<p<0.05$
Articulation - Phone Rate	HAMD Psychomotor Retardation	-0.46	$p = 3.2e-7$
	HAMD Weight Loss	-0.19	$0.01<p<0.05$

Gray shading indicates cases of high significance with $p < 0.01$.

All significant[3] correlations of phone rate with depression ratings are shown in Table I. Examining the HAMD total score, we see that a significant correlation occurs between this total and phone-based speaking rate. The articulation rate measure did not show the same correlation with HAMD total, but did show a stronger relationship with the Psychomotor Retardation rating than the more general speaking rate. The most significant correlations for both speaking and articulation rate measures are with the Psychomotor Retardation ratings. This finding is consistent with the fact that the HAMD Psychomotor Retardation sub-topic is a measure of motor slowing, including the slowing of speech (see Appendix).

Although our rate measurement methods are different, we observe certain consistencies in our findings with those of Mundt et al. [16]. In the Mundt et al. study, on the same database, speaking rate was measured in terms of syllables/second, based on the fixed number of syllables in the "Grandfather Passage". Mundt et al. found a Pearson correlation between HAMD total score and the speaking rate of -0.23 with high significance, consistent with our Spearman correlation of -0.22 for phone-based speaking rate. By computing our measures from the free-response interview section of the recordings, instead of the read-passage recordings, we focus more on the changes in conversational speech and remove the variable of different reading styles used by the patients. In addition, the use of an automatic method allowed us to analyze much longer samples of speech and thus obtain a more reliable estimate.

V. PHONE-SPECIFIC MEASUREMENTS

Up to this point, we have examined global (i.e. average over all phones) measurements of rate across utterances. In this section, we decompose the speech signal into individual phones and study the phone-specific relationships with depression severity. With this approach, we find distinct relationships between phone-specific duration and the severity of certain symptoms, presenting a snapshot of how speech can differ with varying symptom severities. We use two different definitions of phone duration: 1) phone boundaries via an automatic phone recognizer and 2) width of the energy spread around the centroid of a signal [21] within defined phone boundaries. Decomposition into phone-specific measures allows for a more refined analysis of speech timing.

As in Section IV, due to the quantized nature of the rankings, Spearman correlation is used to determine if a monotonic relationship exists between extracted speech features and depression-rating scores.

A. Duration from Phone Recognition Boundaries

Using an automatic phone recognition algorithm [24], we detect the individual phones and their durations. Before proceeding with vowel and consonant phones, we will first examine the silence or 'pause' regions within a free-response speech session.

Pause length: The automatic phone recognition algorithm categorizes pauses as distinct speech units, with lengths determined by estimated boundaries. Both average pause length and percent total pause time are examined in our correlation measures, the results are summarized in Table II.

We compute the correlations between the average pause length over a single speech session and the HAMD total and corresponding sub-topic ratings; the results are shown in Table II. The average pause length is inversely related to the overall speaking rate and so, as seen with the phone-based global speaking rate measures of Section IV, the HAMD Psychomotor Retardation score again shows the highest correlation value. The HAMD total score, along with a large number of sub-topics, show a significant worsening of condition with longer average pause length.

[3] Different categories of significance are given by: $p< 0 .01$ highly significant; $p<0.05$ significant; $p>0.05$ not significant.

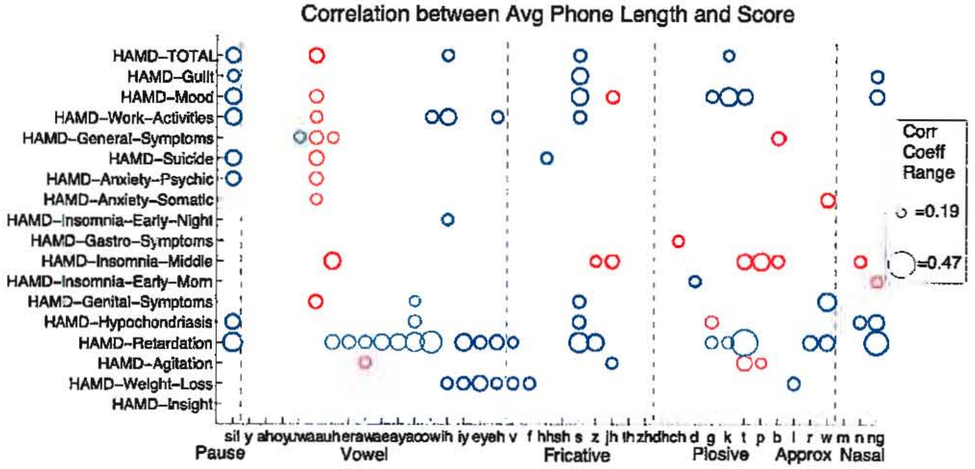

Figure 3: Plot of the correlation between individual phone length and HAMD score. Blue indicates a positive correlation, red a negative correlation. The size of the circle marker is scaled by the magnitude of the correlation. Only significant correlations (p-value < 0.05) are shown. Correlation coefficient range: max marker = 0.47; min marker = 0.19. Correlation results with pause length are included for comparison.

The ratio of pause time measure is defined as the percent of total pause time relative to the total time of the free-response speech session. This feature, in contrast to the average pause length measure, is more sensitive to a difference in the amount of time spent in a pause period, relative to the time in active speech. Thus, a change in time spent thinking, deciding, or delaying further active speech would be captured by the ratio of pause time measure. For this ratio, a highly significant correlation was seen with only the HAMD total score. Most of the significant correlations with total and sub-topic symptom scores seen with ratio of pause time were also correlated with average pause length; the only sub-topic that does not follow this rule is the HAMD measure of Early Morning Insomnia, which shows a higher pause ratio with worsening of condition.

As shown in Table II, we again observe consistency with certain results from Mundt et al. [16] who obtained a Pearson correlation of 0.18 (p-value<0.01) between percent pause time and the HAMD total score, in comparison to our Spearman correlation of 0.25 (p-value = 0.009) between ratio of pause time and the HAMD total score. Mundt also examined a number of pause features for which we do not show results, including total pause time, number of pauses, pause variability, and vocalization/pause ratio. Mundt et al. achieved their highest correlation of 0.38 (p-value<0.001) between the HAMD total score and the pause variability measure. In our own experiments, we did not find a significant correlation between pause variability and HAMD total score. This inconsistency may be due to the difference in speech samples used; we used only the free-response interview data, while Mundt et al. used a variety of speech samples including the free-response, a read passage, counting from 1 to 20, and reciting of the alphabet.

TABLE II
SCORE CORRELATIONS WITH PAUSE FEATURES

Measure	Score Category	Spearman Correlation	p-value
Pause Length	HAMD Mood	0.28	p = 0.003
	HAMD Guilt	0.20	0.01<p<0.05
	HAMD Suicide	0.27	p = 0.004
	HAMD Work and Activities	0.28	p = 0.002
	HAMD Psychomotor Retardation	0.33	p = 0.0003
	HAMD Anxiety Psychic	0.24	p = 0.009
	HAMD Hypochondriasis	0.26	p = 0.005
	HAMD TOTAL	0.26	p = 0.005
Ratio of Pause Time	HAMD Guilt	0.21	0.01<p<0.05
	HAMD Insomnia Early Morning	0.20	0.01<p<0.05
	HAMD Work and Activities	0.19	0.01<p<0.05
	HAMD Anxiety Psychic	0.24	0.01<p<0.05
	HAMD TOTAL	0.25	p = 0.009

Pauses are identified by the phone recognizer; the average of all durations per session is used as the feature. Gray shading indicates cases of high significance with $p < 0.01$.

Phone length: The duration of consonants and vowels, henceforth referred to as *phone length* (in contrast to *pause length*), varied in a non-uniform manner over the observed depression severities. Specifically, the severity of each symptom sub-topic score exhibited different corresponding phone length correlation patterns over all of our recognition-defined phones.

In order to test the correlation between specific phone characteristics and the sub-topic ratings of MDD, average length measures for each unique phone were extracted for each subject and session day. Significant correlations (i.e., correlations with p-value<0.05) across phones are illustrated in Figure 3 for HAMD total and sub-topic ratings. We see that the sign and magnitude of correlation varies for each symptom sub-topic, along with which of the specific phones show significance in their correlation value. A clear picture of the manner of speech (in terms of the phone duration) while certain symptoms are present can be inferred from Figure 3.

The HAMD Psychomotor Retardation correlations stand out across a large set of phones, with positive individual correlations indicating a significant lengthening of these phones with higher Psychomotor Retardation rating. This is again consistent with the slowing of speech being an indicator of psychomotor retardation, but narrows down the phones which are affected to a small group, and reaches the high individual correlation of 0.47 with the average phone length of /t/. In contrast, there are also sub-topics that show groupings of phones that are significantly shortened with worsening of condition; for example, HAMD Insomnia Middle of the Night. Though there is some overlap in the unique phones that show significant correlations with ratings of condition, we see that none of the total or sub-topic correlation patterns contain the exact same set of phones. Nonetheless, strong intercorrelations between the HAMD symptom sub-topics may be seen in the phone correlation patterns; for example, Psychomotor Retardation is most strongly correlated (negatively) with the Agitation subtopic (see Section III); as a possible reflection of this, two phones that show a positive correlation with the Psychomotor Retardation sub-topic are negatively correlated with Agitation. We see that the total HAMD score shows relatively low or no significant correlation values with our individual phone length measures, and the few that do show some significance create a mixed pattern of shortening and lengthening of those phones. Since the total assessment score is composed by taking the sum over all sub-topics, and each sub-topic seems to have a distinct lengthening or shortening speech rate pattern related to it, the total score should only show correlations with phone lengths that have consistent positive or negative correlations across a number of sub-topics; we see that this is the case, especially with pause length (/sil/) and the phones /aa/ and /s/.

An important consideration is the correlation patterns of phones that are produced in a similar way, i.e., have the same manner of articulation. Figure 3 displays the phones in their corresponding groups; dashed vertical lines separate categories (vowel, fricative, plosive, approximant, nasal). We examine each category individually:

Pauses - We include pauses in Figure 3 for comparison. As already noted, longer average pause lengths are measured with worsening of condition for a number of sub-topics (see Table II for correlation values).

Vowels - /aa/ and /uh/ are the two vowels that show more than one significantly negative correlation with a sub-topic, indicating shortening of duration with worsening of condition. There are two groups of vowels that show a positive correlation with HAMD Psychomotor Retardation score 1) the /aw/, /ae/, /ay/, /ao/, and /ow/ group, all which also fall into the phonetic category of open or open-mid vowels, and 2) the /iy/, /ey/, /eh/ group, which also has correlations with the Weight loss sub-topic (in addition to the Psychomotor Retardation sub-topic), this group falls into the phonetic category of close or close-mid vowels.

Fricatives - The fricative which has the most similar correlation pattern to any vowels is /v/, which is a voiced fricative. Consonants /s/ and /z/ both show lengthening (positive correlation) with worsening of Psychomotor Retardation; they are also both high-frequency fricatives. /s/ shows a consistent positive correlation pattern across a range of sub-topics, the correlation pattern for this fricative is most similar to the ones seen for pause length.

Plosives - With regard to Psychomotor Retardation, the three plosives which show significant positive correlations are /g/, /k/, and /t/, which are also all mid to high-frequency plosives; this group also shares similar correlations for the Mood sub-topic. A smaller effect, /t/ /p/ and /b/, all of which are diffuse (created at the front of the mouth, i.e., labial and front lingual) consonants, all show negative correlations with Middle of the Night Insomnia.

Approximants - Both /r/ and /w/ show a positive correlation with Psychomotor Retardation. The single significant correlation found for /l/ is with the Weight Loss sub-topic, which has no other correlation within the approximant group, but does show consistent correlations with a subset of the vowel (/ih/, /iy/, /ey/, /eh/) and fricative (/v/, /f/) groups.

Nasals - The nasal /m/ had no significant correlations with HAMD rating. The nasal /n/ has two significant correlations, but does not have similar correlation patterns to any other phone. The phone /ng/ has a correlation pattern most similar to /s/ and pauses.

We provide additional analysis of the correlation patterns across phones, with respect to the intercorrelations between HAMD sub-topics, in our Conclusions (Section VII).

As an extension of the individual phone results, sub-topics with at least four significant individual phone correlations were identified and linearly combined into a measure. Positive or negative unit weights were chosen based on the sign of their individual phone correlation values. More formally, denote the average length of phone k by L_k and suppose a subset P_i is the set of significantly correlated average phone lengths for HAMD sub-topic i. We then define a new variable L^i as the sign-weighted sum

$$L^i = \sum_k \alpha_k L_k \qquad k \ni P_i$$

where the weighting coefficients α_k are ±1, defined by the sign of the relevant phone correlation. The full feature extraction process, from speech to the final linearly-combined duration measure, is outlined in Figure 4.

Through this simple linear combination of a few phone-specific length features we are able to achieve much higher correlations than when examining average measures of the speech (i.e., globally), and, as before, the highest correlation is reached by the HAMD Psychomotor Retardation sub-topic.

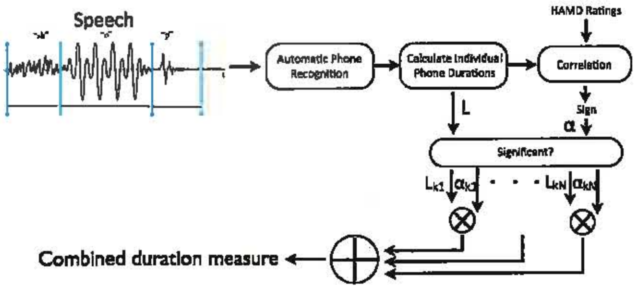

Figure 4: Overview of the method for computing the combined duration measure. For this example, there is a subset of *N* significant phone duration correlations, indicated by *k=k1 ... kN*.

The resulting correlation between the weighted sum of the individual phone lengths and the relevant score is shown in Table III. The left-most column gives the set of phones used for each sub-topic (selected based on correlation significance). We observe that our largest correlations thus far are reached by our "optimally" selected composite phone lengths with each sub-topic. The largest correlation of the composite phone lengths is again reached by the HAMD Psychomotor Retardation measure with a value of 0.58, although the gain in correlation value from 0.47 (achieved with /t/) to 0.58 is small considering the large number of phones that contribute to the composite feature (19 phone durations and pause/silence duration). In contrast, for the HAMD Work and Activities sub-topic, a correlation gain from 0.28 (/ih/) to 0.39 (/sil/ /aa/ /ih/ /ow/ /eh/ /s/) is achieved using only 6 phone lengths in the composite feature.

TABLE III
SCORE CORRELATIONS WITH SIGNED AGGREGATE PHONE LENGTH

Phones used	*Score Category*	*Spearman Correlation*	*p-value*
(sil, aa, g, jh, k, ng, s, t)	HAMD Mood	0.43	p=2.7e-6
(uh, b, jh, n, p, t, z)	HAMD Insomnia Middle of the Night	0.37	p=6.8e-5
(sil, aa, ih, ow, eh, s)	HAMD Work and Activities	0.39	p=2.7e-5
(sil, ae, iy, ay, ey, ao, ow, eh, aw, uh, er, g, k, ng, r, s, t, v, w, z)	HAMD Psychomotor Retardation	0.58	p=1.7e-11
(aw, jh, p, t)	HAMD Agitation	0.34	p=2.0e-4
(aa, uw, uh, b)	HAMD General Symptoms	0.40	p=1.4e-5
(aa, ao, s, w)	HAMD Genital Symptoms	0.42	p=4.5e-6
(sil, ao, g, n, ng, s)	HAMD Hypochondriasis	0.39	p=2.0e-5
(iy, ey, ih, eh, f, l, v)	HAMD Weight Loss	0.39	p=2.6e-5
(sil, s, k, ih, aa)	HAMD TOTAL	0.35	p=1.8e-4

An alternative view of the correlation results of Table III is shown in Figure 5. Here we display a comparison between the highest individual phone correlation and the composite length feature correlation values from Table III. Significant correlations with global speaking rate (from Table I) are included for comparison.

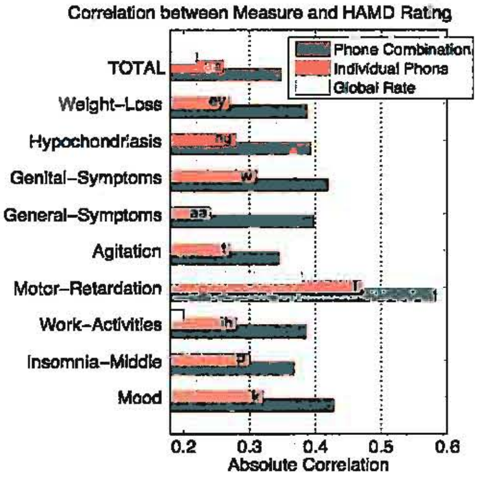

Figure 5: Absolute Spearman correlation value between measure and HAMD score. The individual phone correlation bars correspond to the maximum absolute correlation between depression assessment score and a single phone-specific average length; the specific phone used is shown at each bar. The phone combination correlation bars show the absolute correlation value between assessment score and the signed aggregate phone length; the phones used for this aggregate length are listed in the first column of Table III. Global speaking rate correlation values from Table I are included for comparison.

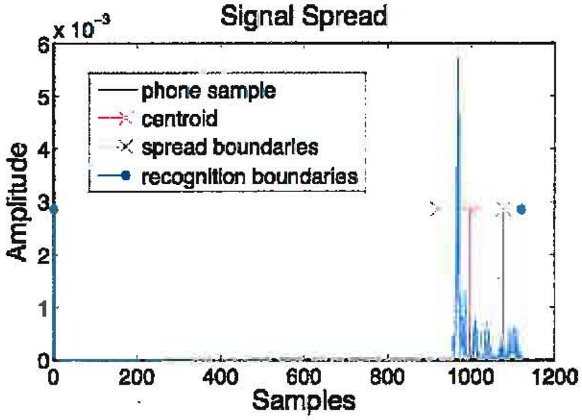

Figure 6: Example of a single utterance of the burst consonant /t/ where the boundaries detected by the automatic phone recognizer are greater than the phone duration corresponding to energy spread. Asterisk and cross markers show our estimated centroid and spread boundaries for this phone.

B. Phone-specific spread measurement

An alternative definition of phone duration was constructed using the concept of the spread of a signal's energy. A large subset of our phones consist of a single, continuous release of energy with tapered onset and offsets, particularly the case with burst consonants (e.g., /p/, /b/, etc.) and vowel onsets and offsets. (See Figure 6 for an example.) In these cases phone boundaries, as deduced from an automatic phone recognizer, may not provide an appropriate measure of phone duration.

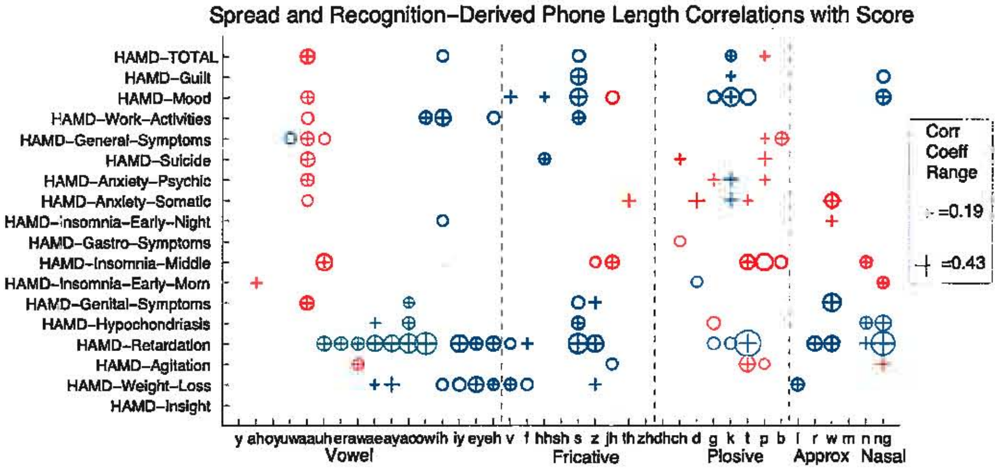

Figure 7: A comparison between the spread and recognition-derived length correlations with depression rating. Spread correlations are marked with a cross, recognition-based length correlations are marked with a circle. Blue indicates a positive correlation, red a negative one. The size of the marker is scaled by the magnitude of the correlation. Only significant correlations (p-value<0.05) are shown. Correlation coefficient range: max cross marker = 0.43; min cross marker = 0.19. Range of circles is the same as in Figure 3.

One measure of phone length or duration is given by the signal spread about the centroid of the envelope of a signal [21]. The centroid of the phone utterance, denoted $e[n]$, is computed via a weighted sum of the signal. Specifically, the centroid for each phone utterance, n_{centroid}, is given by

$$n_{centroid} = \sum_{n=1}^{N} n \frac{e[n]^2}{\sum_{m=1}^{N} e[m]^2}$$

where the square of the signal is normalized to have unit energy and where N is the number of samples in each phone utterance. The standard deviation about $n_{centroid}$ is used as the 'spread' (i.e., alternate duration) feature. The spread of a single phone utterance is thus calculated as

$$spread = \sqrt{\sum_{n=1}^{N} (n - n_{centroid})^2 \frac{e[n]^2}{\sum_{m=1}^{N} e[m]^2}}$$

Significant spread-based phone length correlations are illustrated in Figure 7 for both HAMD total and sub-topic ratings. We see again that HAMD Psychomotor Retardation stands out with a large set of significant positive correlations with phone duration, indicating longer durations with worsening of condition. HAMD Insomnia Middle of the Night shows consistent shortening of phone duration with increasing severity ratings. This consistency with the recognition-based length results is a product of the strong correlation between our recognition and spread-based measures. We see that overall there are *more changes in the correlation results with burst consonants*, such as /k/, /g/, and /p/, than with any other phones due to their burst-like, shorter nature in time. As seen in Figure 6, the phone recognition algorithm showed a tendency to overestimate (set too early) the onset phone boundary for these burst consonants; on the other hand, the duration of the silence gap prior to or after the burst may also be condition-dependent.

C. *Effects of Noise and Sub-topic Intercorrelation*

One of the more general relationships that can be drawn from this data is that worsening of psychomotor retardation condition can be observed in a subject's speech rate. A question we can then ask is "Are the correlations between our speech measures and the other sub-topics the result of noise and/or sub-topic intercorrelation with the Psychomotor Retardation sub-topic?" In order to alleviate the effects of spurious correlations on our interpretation, in addition to only showing significant results, the presentation of the results in Figures 3, 7, 11 and 12 is such that phones are grouped according to manner of articulation and the sub-topics are grouped by significant absolute intercorrelation values. Clustering of significant correlations within a phonetic or intercorrelation sub-group suggests that these consistent correlations are indeed meaningful.

For further applications, one needs to know which correlation results are the product of strong intercorrelation between each sub-topic and Psychomotor Retardation and which are not. To help address this, although this likely deserves a more in-depth analysis, an additional experiment was run where the correlations between sub-topics and phone length were re-computed using only the speaker-session samples that had a Psychomotor Retardation score of 0 (i.e., no recorded psychomotor retardation). The results can be seen in Figure 8, we see that for sub-topics that are strongly correlated with Psychomotor Retardation, such as Agitation and Work-Activities (see Figure 2), the correlation patterns do change and most of the significant correlations found earlier are no longer present. For sub-topics that have a weak correlation to Psychomotor Retardation, such as Suicide or General Symptoms, we see that many of the previous significant correlations found with phone length remain the same. In addition, we see that for all correlations that are retained with this second analysis there is not a change in sign, further supporting the hypothesis that these correlations are not spurious or completely due to intercorrelations with Psychomotor Retardation.

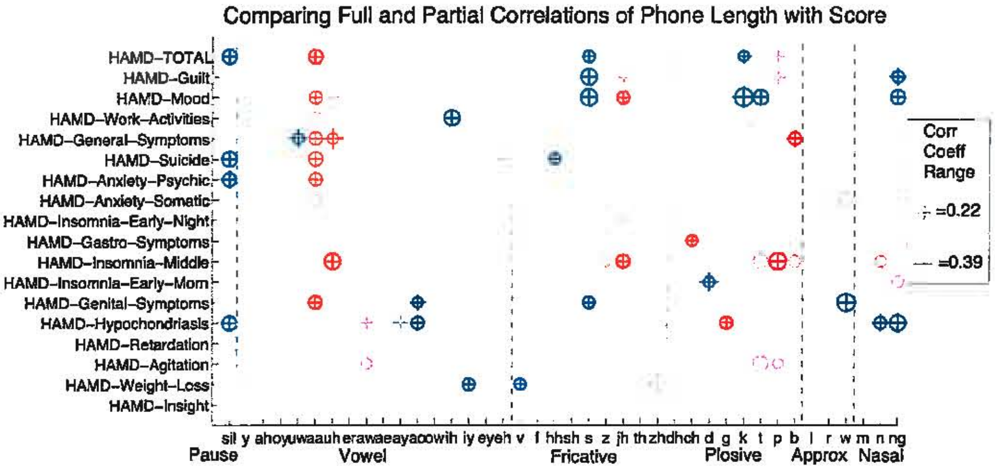

Figure 8: A comparison between the individual phone length correlations of Figure 3 (circle marker) and the individual phone length correlations when all samples showing a nonzero Psychomotor Retardation rating are removed from the calculation (cross marker). Correlations which are not significant in both cases are faded for visualization. Blue indicates a positive correlation, red a negative one. The size of the marker is scaled by the magnitude of the correlation. Only significant correlations (p-value<0.05) are shown. Correlation coefficient range: max cross marker = 0.39; min cross marker = 0.22.

D. Phone Recognition Accuracy

As mentioned earlier, the phone recognition algorithm is based on a Hidden Markov Model approach, which for English was reported as having about an 80% overall accuracy [24]. Although this implies some mislabeling of phones, the mislabeling is often between similarly structured (i.e. similar in time and frequency) phones. The primary effect of labeling errors is a form of added 'noise' to our correlation studies and the feature vectors in Sections V and VI. In spite of this noise inclusion, we found strong correlations with phone-specific length features and support these feature results with the preliminary classification work of Section VI. Nevertheless, a more quantitative study of the effect of phone mislabeling is warranted.

VI. CLASSIFIERS OF MDD: PRELIMINARY RESULTS

Our correlation results motivate the development of automatic classifiers of depression severity based on phone-specific measures of speech rate. Feedback from a reliable classifier would be a highly beneficial tool for clinicians. Reliable classifiers could even be used as a tool to aid in the standardization of depression ratings. As an initial step toward this goal, we provide a proof-of-concept use of speech rate features, specifically, the set of recognition-derived, phone-specific lengths, for classification. A more exhaustive classification study requires a larger, more comprehensive database and investigation of the broader suite of speech rate features, such as the phone length from energy spread or signal power; we speak to this in our on-going work, Section VII.

In forming depression classifiers, we consider the 5-class problem for the HAMD total score; the 5-class case is divided into the ranges 0-5, 6-10, 11-15, 16-20, and 21-27. A 5-class experiment demonstrates a test of classification accuracy. For the symptom sub-topics, we implemented the 3, 4, or 5-class problem for each sub-topic based on the maximum possible range for each; for example, the HAMD Mood sub-topic has the possible scores of 0, 1, 2, 3 or 4, thus we implemented a 5-class problem for this sub-topic. For all classifiers considered, we test using a leave-one-out cross validation scheme, as illustrated schematically for a 2-class case in Figure 9.

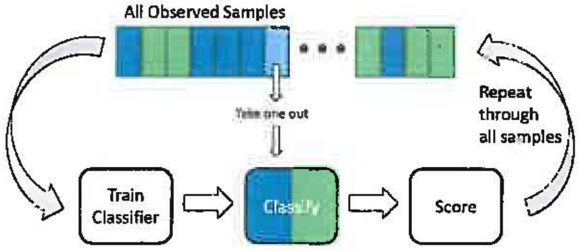

Figure 9: Illustration of the leave-one-out cross-validation approach for the 2-class problem, depicted as green vs. blue. Each unique subject-session pair in our dataset is an 'observed sample' that is described by its feature vector. For cross validation, we take one sample out, train the classifier on the remaining samples, classify the excluded sample and record the performance. The process is repeated until all of the observed samples have been tested.

We use a simple Gaussian maximum-likelihood algorithm for all experiments; i.e., each class is modeled as a multi-dimensional Gaussian, with the number of dimensions matching the feature vector dimension, classification is then performed by finding the class of maximum likelihood for the test sample. Our phonological feature vector is composed from our recognition-derived average phone (vowels and consonants) lengths (see Section V.A) and the average pause (silence) length values. We consider four different feature selection methods: 1) A single feature, the signed aggregate of individual phone lengths and pause length - see Table III, column 1 for a selection of phones used (Signed Agg); 2) No feature selection, i.e., use all individual average phone lengths and/or the pause length as a vector of features (None); 3)

Hand-selection of the subset of individual phone lengths and/or pause length that show significant correlation statistics to form a feature vector (Stat Sig); and 4) A subset of individual phone lengths and/or the pause length is automatically selected to minimize error, though an optimal solution is not guaranteed (Min Error) [11].

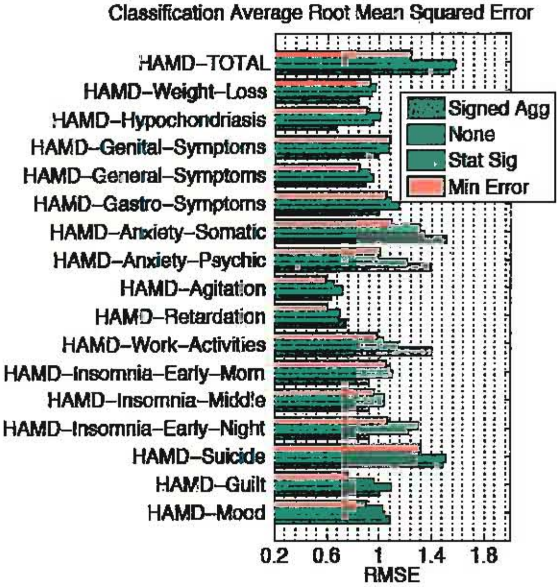

Figure 10: Adjusted Root Mean Squared Error for classification in the 3, 4, or 5-class case (depending on the range of possible ratings) for symptom sub-topics, and the 5-class case for total HAMD score. Different color bars indicate the method of feature selection, 'none' being no feature selection (i.e. all phone length features used). See text for description of the four feature sets.

Providing classification results on the symptom sub-topics would add an additional level of feedback to a clinician. In addition, considering each rating level as a class takes into account the fact that variations on a single-point scale could indicate large changes in an individual's condition. We therefore examine each sub-topic as a 3, 4, or 5-class problem, with the number of classes matching the range of possible scores for each particular sub-topic. We also divide the total scores into a 5-class problem in order to test the classifier's ability to differentiate between in remission, mild, moderate, severe, or very severe depression. We found that most of the classification errors come from misclassification into an adjacent severity level, for example a severity rating of 1 for a given sub-topic might be misclassified as a 0 or a 2. These results are summarized in Figure 10, which shows the (average-adjusted)[4] root mean squared error (RMSE) for each individual assessment rating. The RMSE gives a sense of how far the classifier diverges from the clinician or self-reported rating; all of the RMSEs fall below 2, quantifying our observation that most misclassifications fall into an adjacent severity level. In almost all cases, we benefit from some form of feature reduction; features that were hand-selected from the correlation results overlap but do not exactly match the features that are chosen by the algorithm to minimize error. Finally, the RMSEs indicate the predictive potential of our phonologically-based feature sets including the single feature of linearly combined duration.

As we are using only a subset of our speech rate features, the recognition-derived average phone lengths and the average pause length, one could potentially improve performance by extending the feature space beyond what is used in this preliminary study. Specifically, we have not used signal power and spread-based features, not to mention other phonetic-based features. Further discussion of such extensions is given in the final section VII.

VII. Conclusions and On-Going Work

A. Conclusions

Our correlation studies give direction in determining which speech-rate-based vocal features may be useful for detecting depression symptoms. For all of our cases, a phone-specific approach showed higher correlations than the global rate measurements. We considered pause length separately from vowel/consonant length due to the different factors that can affect the two types of features; we assume that pause length incorporates both psychomotor issues along with possible hesitancy due to other depression symptoms. The usage of energy spread to define phone duration provides an alternate scheme for computing phone duration, not tied strictly to automatic recognition-based phone boundary definitions. The phone and symptom-specific correlation patterns present a visual interpretation of how speech can change with different symptom severities. Possibly, speech sounds with either similar production categories or similar usages in speech (e.g., at the onset or at the ending of a word) would show correspondingly similar changes with MDD condition severity; we explored the former by grouping the phones by manner of articulation and finding consistencies in the correlations within the groups. Other experiments that indicated not all meaningful sub-topic correlations are tied to Psychomotor Retardation involved correlations between sub-topics and phone length re-computed using only the speaker-session samples that had a Psychomotor Retardation score of 0. The additional correlation study with the linearly combined phone duration measure shows how using only a subset of phones can reveal a stronger underlying relationship.

[4] Adjusted Root Mean Squared Error (RMSE) is the average over the RMSE for each rating value, giving each an equal weight, to account for some highly skewed distributions of observed data.

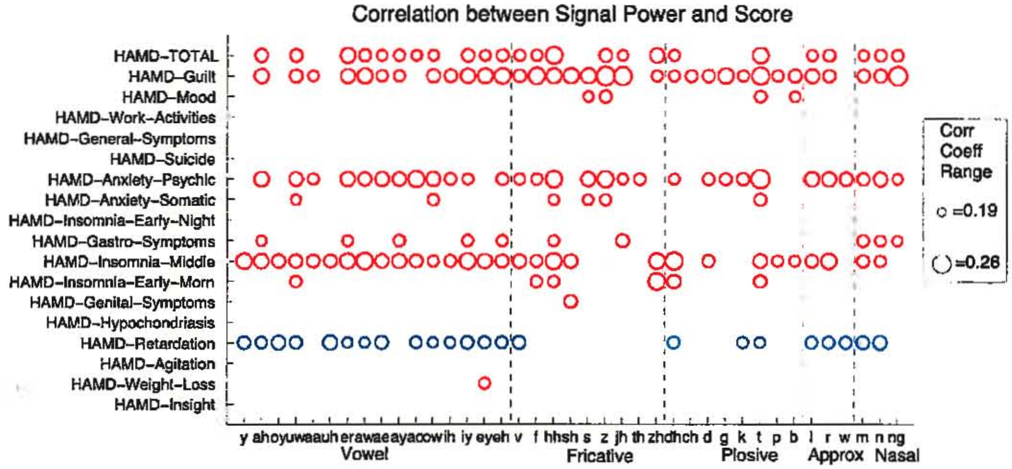

Figure 11: Plot of the correlation between average phone power and HAMD score. Blue indicates a positive correlation, red a negative one. The size of the circle is scaled by the magnitude of the correlation. Only significant correlations (p-value<0.05) are shown. Correlation coefficient range: max marker = 0.26; min marker = 0.19.

Our correlation results show a snapshot of how speech can vary across each individual symptom severity. Another possibility that we considered is that sub-topics with similar or correlated symptoms would show similarities in the shift in speech rate and phone-specific duration measures. The similarities between symptom sub-topics are quantified by the intercorrelations shown in Figure 2. As an example analysis, we examine the Psychomotor Retardation sub-topic which is most strongly correlated with Agitation (negatively, -0.40) and Mood (positively, 0.36). Keeping this in mind, we see in Figure 3 oppositely-signed significant correlations for both Psychomotor Retardation and Agitation for 2 phones (/aw/, /t/); we also see positive significant correlations for both Psychomotor Retardation and Mood for the same 5 phones and the pause measure (/sil/, /s/, /g/, /k/, /t/, /ng/). The strongest HAMD intercorrelation for our dataset falls at 0.64 and corresponds to the correlation between the Mood and Work-Activities sub-topics. Although the correlation patterns for these phonologically-based measures share some characteristics, they are not the same, indicating that the two sub-topics are somewhat distinct.

We have also introduced a preliminary study for classification of depression severity based on our speech-rate features using phone length derived from phone-recognition boundaries. Using a simple Gaussian-likelihood classifier, we show the results for the 3, 4, or 5-class classification problem for all HAMD score categories, with each class representing a different severity level. Our preliminary classification results show promise as a beneficial tool to the clinician; both as an initial measure of depression level and in assessing severity of symptoms, and motivate the extension of the work to further phone-based features.

Depression does not have the same symptom progression in all patients and should not be treated as such. Our correlation and classification results with the HAMD MDD assessment reveals changes that occur in speech rate with different symptom severities. Some symptoms, such as Psychomotor Retardation, have a consistent relationship with a change in speech pattern, while others, such as short-term changes in Weight, may not. Identifying reliable biomarkers for each symptom is useful, since each symptom category and progression to different severities is more homogeneous across patients than the overall depression rating, which can encompass completely different manifestations of the disorder.

In this paper, we found significant correlations between a subset of the HAMD symptom sub-topic ratings and our vocal features, with supporting classification results. We found that a symptom-specific approach offers a more informative profile of a subject's state and is more likely to result in consistent shifts in speech pattern or behavior. In the case of the total HAMD score, however, the case-by-case variability with which different sub-topics will increase in severity with worsening of MDD condition and the sub-topic-specific relationships that we see with speech measures suggests that one might not be able to expect a high HAMD total score to coexist with a reliable shift in a particular speech pattern. Each symptom sub-topic, when examined individually across its entire severity range, has unique and sometimes opposing shifts in speech rate measures.

B. On-going Work

Based on the success of phone-specific speech rate measures in correlating with certain MDD symptoms, we plan to extend our experiments to examining other phone-specific speech measures, thus exploiting the general phonological framework that we have developed. Our continuing studies include phone-specific energy measures, an examination of vowel usage in depression, and measures involving prosodic rhythm and modulation [23], and using the derivative of measures. The derivative of a vocal feature allows one to track how the changes in an individual's speech pattern may match similarly scaled changes in their condition. Use of derivatives also serves as a way to normalize out absolute levels in a subject's baseline speech.

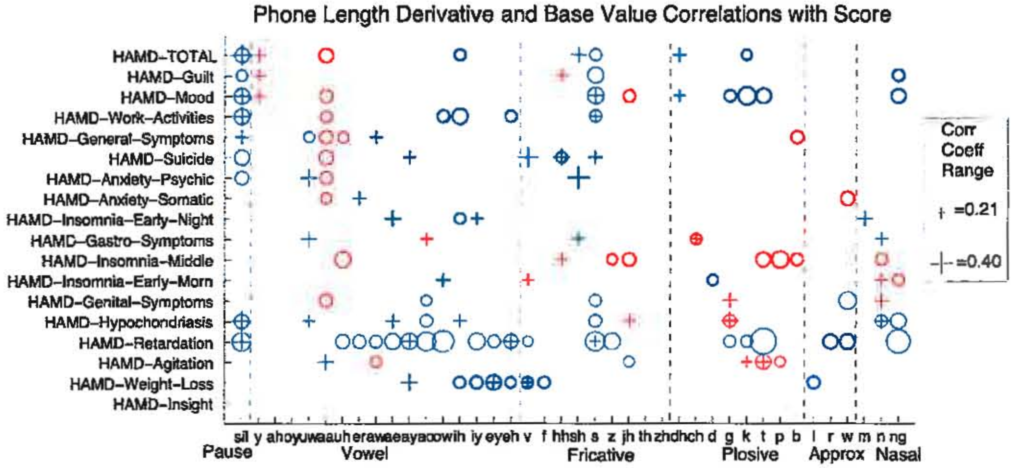

Figure 12: A comparison between the phone length derivative and base value (see Section V.A) correlations with depression rating. Derivative correlations are marked with a cross, base value correlations are marked with a circle. Blue indicates a positive correlation, red a negative one. The size of the marker is scaled by the magnitude of the correlation. Only significant correlations (p-value<0.05) are shown. Correlation coefficient range: max cross marker = 0.40; min cross marker = 0.21.

As a taste of our on-going work, we cover a series of phone-based measures that extend the present results. We first discuss an alternative speech unit for computing speech rate, the *pseudo-syllable rate*. Individual phones are combined such that each vowel forms the nucleus of its own segment, with all of the proceeding consonants grouped with it. Thus, a measure of pseudo-syllable rate will be highly correlated to the phone rate results. The motivation for this unit is its relation to syllables and the difficulty in automatically extracting syllables [22]. The speaking and articulation rate, as defined in Section IV, were calculated with respect to the pseudo-syllable rate and correlations with HAMD scores were computed. Similar to the phone rate results, the pseudo-syllable speaking rate shows significant correlation with the HAMD Psychomotor Retardation (-0.37) and total (-0.26), and the pseudo-syllable articulation rate shows highly significant correlation with the Psychomotor Retardation rating (-0.41).

Continuing our phone-based measures, we show in Figure 11 a correlation plot for individual phone average power. Phone power is computed as the sum of the squared signal over time. We see that the significant correlations with phone power are more uniform across phones within a sub-topic. Correlations with Psychomotor Retardation are negative for all phones and limited to mostly the vowel, approximant and nasal phone categories.

In Figure 12, we show a plot comparing the individual phone length correlations of Figure 3 to the corresponding derivatives of the phone lengths. A rough derivative of the vocal features was computed by measuring the relative change between feature values on consecutive session days for each subject. The corresponding derivative of the depression ratings was computed in the same way. Comparing the derivatives results with the base value phone-specific correlations, there are no inconsistencies in the direction of length change with severity of condition; in other words, for all overlapping significant correlations, no positive correlation in one study is negative in the other.

In this paper, we have only touched on classification-algorithm development, illustrating the predictive potential of our phonologically-based features including the single feature of a simple linearly-combined phone duration. We plan to extend this preliminary study using both more sophisticated classification schemes, such as the use of Support Vector Machines (SVMs) and a more comprehensive set of speech features such as variations of our speech-rate measures, power, fundamental frequency measures, and temporal- and frequency-based rhythmic/modulation patterns. Along these lines, we will draw on prosodic tokenization approaches applied in other contexts [22][23].

We touch on the issue of automatic phone recognition errors that can affect the accuracy of our speech-rate measures (see Section V.D). We plan to further investigate the effect of these errors on our correlation and classification results. For example, the current phone recognizer [24] might be improved by invoking utterance transcriptions. Finally, we plan to explore the complementary use of other joint modalities, such as video tracking of facial features (e.g., visemes), that can yield biomarkers for certain symptoms or mental conditions that do not necessarily show in speech patterns.

More generally, we suspect that for other types of vocal features besides speech rate, the phone-specific approach, along with an individual MDD symptom analysis, will result in a more accurate representation of how speech can vary with different progressions of MDD.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. James Mundt for many helpful discussions on details and advice on the use of his data collection, as well as providing his invaluable knowledge and experience on Major Depressive Disorder. The authors also thank Dr. Pedros Torres-Carrasquillo, Dr. Douglas Sturim, Fred Richardson, and Dr. Michael Brandstein from MIT Lincoln Laboratory for assistance in use of classification algorithms. Finally, the National Institute of Mental Health supported Dr. James Mundt in the data collection used in this paper. Finally, the authors acknowledge the paper's editor and reviewers for detailed and insightful comments that helped improve the manuscript.

APPENDIX: CLINICAL HAMD ASSESSMENT COMPONENTS

HAMD sub-topics: The range of score for each is included in parenthesis; higher scores indicate a worsening of condition.

1) Depressed Mood (0-4) – Sadness, hopeless, helpless, worthless, along with the person's inability to hide these feelings
2) Feelings of Guilt (0-4) – Magnitude of guilt
3) Suicide (0-4) – Thoughts of suicide along with severity of attempts
4) Insomnia: Early in the Night (0-2) – Difficulty falling asleep
5) Insomnia: Middle of the Night (0-2) – Waking during the night
6) Insomnia: Early Hours of the Morning (0-2) – Early waking and inability to return to sleep
7) Work and Activities (0-4) – Thoughts or feelings of fatigue and level of interest in work or activities
8) Psychomotor Retardation (0-4) – Slowness of thought and speech, impaired ability to concentrate, decreased motor activity
9) Agitation (0-4) – Physical inability to sit still
10) Anxiety Psychic (0-4) – Level of expression of anxiety
11) Anxiety Somatic (0-4) – Physiological concomitants of anxiety
12) Somatic Symptoms Gastro-intestinal (0-2) – Loss of appetite, heavy feeling in abdomen
13) General Somatic Symptoms (0-2) – Heavy limbs, muscle aches, headache, fatigue
14) Genital Symptoms (0-2) – Loss of libido, menstrual disturbances (for women)
15) Hypochondriasis (0-4) – Magnitude of hypochondria
16) Loss of weight (0-3) – Magnitude of weight loss in previous week
17) Insight (0-2) – Denial of illness

REFERENCES

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision. Washington, DC, American Psychiatric Association, 2000.

[2] Bagby, R. M., A. G. Ryder, M.A., D. R. Schuller, and M. B. Marshall, "The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight?", Am J Psychiatry 161:2163-2177, December 2004.

[3] Cannizzaro, M., B. Harel, et al. (2004). "Voice acoustical measurement of the severity of major depression." *Brain and cognition* 56(1): 30-35.

[4] Cohn, J., T. Kruez, et al. (2009). "Detecting depression from facial actions and vocal prosody." *Emotion* 10: 18-19.

[5] Emotion Challenge, *Proceedings of Interspeech 2009*, Brighton, UK.

[6] Fava, M. and K. Kendler (2000). "Major depressive disorder." *Neuron* 28(2): 335-341.

[7] Flint, A., S. Black, et al. (1993). "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression." *Journal of psychiatric research* 27(3): 309-319.

[8] France, D., R. Shiavi, et al. (2000). "Acoustical properties of speech as indicators of depression and suicidal risk." *IEEE Transactions on Biomedical Engineering* 47(7): 829.

[9] Hamilton, M. (1960). "A rating scale for depression." *British Medical Journal* 23(1): 56.

[10] Lemke, M., P. Puhl, et al. (1999). "Psychomotor retardation and anhedonia in depression." *Acta Psychiatrica Scandinavica* 99(4): 252-256.

[11] Kukolich, L., and Lippman, R., LNKnet, MIT Lincoln Laboratory, February, 2004.

[12] Low, L.A., Maddage, Lech, M., Sheeber, L., Allen, N. (2010). "Influence of acoustic low-level descriptors in the detection of clinical depression in adults," *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

[13] Mendenhall, W., R. Beaver, et al. (2008). Introduction to probability and statistics, Brooks/Cole.

[14] Moore, E., M. Clements, et al. (2008). "Critical analysis of the impact of glottal features in the classification of clinical depression in speech." *IEEE Transactions on Biomedical Engineering* 55(1): 96-107.

[15] Moore II, E., M. Clements, et al. (2003). "Analysis of prosodic variation in speech for clinical depression." *Proceedings of the 25th Annual International Conference of the IEEE EMBS*: 2925-2928.

[16] Mundt, J., P. Snyder, et al. (2007). "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology." *Journal of Neurolinguistics* 20(1): 50-64.

[17] Myers, J. and A. Well (2003). *Research design and statistical analysis*, Lawrence Erlbaum.

[18] Ozdas, A., R. Shiavi, et al. (2004). "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk." *IEEE Transactions on Biomedical Engineering* 51(9).

[19] Ozdas, A., R. Shiavi, et al. (2004). "Analysis of vocal tract characteristics for near-term suicidal risk assessment." *Changes* 2: 3.

[20] Pittam 1993 Handbook of Emotions, Ch. 13.

[21] Quatieri, T.F., *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2001.

[22] Rouas J., "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination, *IEEE Trans. Audio, Speech, and Language Proc.*, Vol. 15, Nop. 6, August 2007.

[23] Rouas J., Farinas J., Pellegrino F., Andre´-Obrech R., "Rhythmic unit extraction and modeling for automatic language identification," *Speech Communication* 47 (2005) 436–456,

[24] Shen, W., White. C, Hazen, T.J., "A comparison of query-by-example methods for spoken term detection,", ICASSP10.

[25] Sobin, C. and H. Sackeim (1997). "Psychomotor symptoms of depression." *American Journal of Psychiatry* 154(1): 4.

[26] World Health Organization (2001). The World Health Report : 2001 : Mental health : new understanding, new hope. Geneva, World Health Organization.